Unsupervised Learning: Introduction

About the Instructors

Genevera Allen:

- Rice University - Departments of Electrical and Computer Engineering, Stat and CS & Baylor College of Medicine - Neurological Research Institute.
- Founder & Director, Rice D2K Lab
- Research:
  - Statistical Machine Learning, Data Integration, Graphical Models, Modern Multivariate Analysis, Interpretability & Fairness, Neuroscience, Genomics.

`https://d2k.rice.edu/`

About the Instructors

Yufeng Liu:

- University of North Carolina, Chapel Hill - Departments of Statistics and Operations Research, Genetics, & Biostatistics.
- Research:
  - Statistical Machine Learning and Data Mining; High-dimensional Data Analysis; Nonparametric Statistics and Functional Estimation; Bioinformatics; Design and Analysis of Experiments.

`http://yfliu.web.unc.edu`

Teaching Assistants

Hui Shen:

- PhD Candidate, Statistics - University of North Carolina, Chapel Hill.

Camille Little:

- PhD Candidate, Electrical & Computer Engineering, Rice University.

Statistical Machine Learning

- "Learn" from current data to make predictions about the future.

  Examples?

- Intersection of: Computer Science, Statistics, Applied Math.

Big Data

Big Data - BIG in Volume, Variety and/or Velocity (or Complexity!).

Common Big Data themes in Statistical Learning:

- Big n. Large number of observations.
  - Examples: Internet data, financial transactions, climate data, etc.
- Big p. Large number of features relative to observations. (High-dimensional data).
  - Examples: Medical data - genomics, neuroimaging, medical imaging, etc.

Big Biomedical Data

Examples:

- High-throughput Genomics ("Omics").
  - RNA-sequencing, microarrays, methylation arrays, CGH-arrays, exome sequencing, mass spectrometry, NMR spectroscopy, etc.
- Neuroimaging / neural recordings.
  - MRI, Functional MRI (fMRI), EEG, MEG, DTI, ECoG, PET, etc.
- Electronic Health Records.
- Medical Imaging.
- Text Data - Pubmed abstracts.

Data Matrix

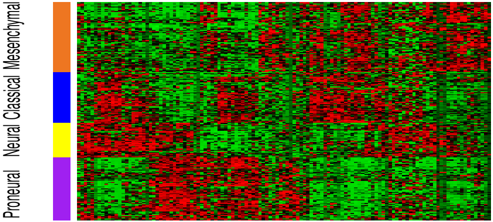
Data Matrix:

$$\boldsymbol{X}_{n\times p} = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1p} \\ \vdots & & \ddots & \\ x_{n1} & x_{n2} & \dots & x_{np} \end{pmatrix}$$

- Rows: n observations / samples / subjects.
- Columns: p features / variables.

Data Matrix

Example: Omics Data



Gene Expression Data (Microarray)

- Rows (observations): Subjects ($n \approx 100 - 500$).
- Columns (features): Genes ($p \approx 500 - 20,000$).
- Measurement: Gene expression levels (loosely, how much a gene is turned off or on in a sample).

Data Matrix

Example: Text Mining

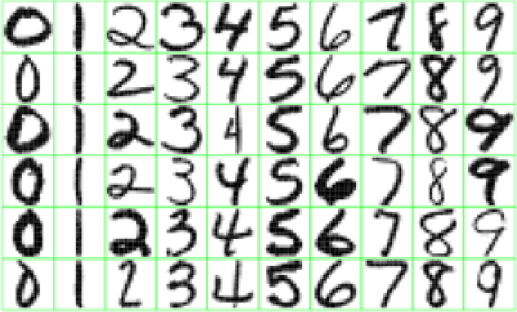
	data	R	big	cluster	shiny	fast	plot
doc 1	57	1	43	2	0	22	4
doc 2	17	29	2	3	35	6	44
doc 3	47	33	0	0	24	3	19
doc 4	23	0	0	31	0	7	2
doc 5	40	5	28	9	0	21	6
doc 6	8	10	7	46	12	17	9

(Bag-of-Words Format)

- Rows (observations): Documents ($n \approx 500 - 100,000$).
- Columns (features): Words ($n \approx 100 - 50,000$).
- Measurement: Count of how many times words appeared in documents.

Data Matrix

Example: Image Data



(Handwritten Digits Data)

- Rows (observations): Digits ($n \approx 10,000$).
- Columns (features): Pixels ($p = 256$).
  - Each digit image is converted to a 16×16 grayscale image. The 256 total pixels are vectorized to form the features.
- Measurement: Normalized grayscale intensity of each pixel.

Unsupervised vs. Supervised Learning

$$\boldsymbol{X}_{n \times p} = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1p} \\ \vdots & & \ddots & \\ x_{n1} & x_{n2} & \dots & x_{np} \end{pmatrix}$$

- Rows: n observations / samples / subjects.
- Columns: p features / variables.

Supervised Learning:

$$\mathbf{y} = (y_1, y_2, \dots y_n)^T$$

- $\mathbf{y}$ - n labels / outcomes associated with each observation.

Unsupervised Learning: No outcomes / labels!

Supervised Learning

Main Goal

Prediction!

- Given: $(Y_n^{train}, \boldsymbol{X}_{n \times p}^{train})$ (Training Data).
- Training: Use training data to find $\hat{f}()$ that maps $\boldsymbol{X}$ to Y: $Y = f(\boldsymbol{X}) + \epsilon$.
- Prediction: Given new $\boldsymbol{X}_{m \times p}^{test}$, predict $Y_{m \times 1}^{test}$: $\hat{Y}^{test} = \hat{f}(\boldsymbol{X}^{test})$.

Examples?

Secondary Goals:

- Feature Selection - What features are associated with the outcome?
- Others?

Unsupervised Learning

No labels! What is the goal?

Main Goal

Find some structure that characterizes the data.

(Or, find structure in training data that we expect to be present in future data.)

- Find patterns. (PCA, ICA, NMF, MDS)
- Dimension reduction. (PCA)
- Group observations / Group features / Group both. (Clustering)
- Find associations / relationships between features or observations. (Graphical or Network Models)
- Filter features. (Association testing)

Unsupervised Learning

Challenges:

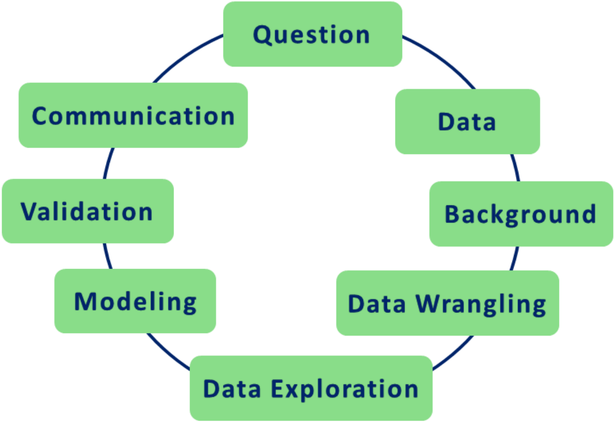
- Difficult to validate unsupervised learning results.
- No validation or test labels to measure prediction accuracy.
- What is meaningful structure in data?

Uses:

- Data pre-processing / compression / denoising.
- Exploratory data analysis.
  - Need to use multiple unsupervised learning techniques as each gives slightly different "insights" into data.
- Data visualization.
- Data-Driven Discovery.

Unsupervised Learning

How does it fit into a data science pipeline?



Unsupervised Learning

How is it used in Big Biomedical Data?

Case Study: BRCA gene expression data.

- Data Visualization.
  - Cluster heatmap, graphical models, MDS, PCA.
- Exploratory Analysis.
  - Clustering / dimension reduction to find cancer subtypes.
- Gene Selection.
  - Large-scale hypothesis testing to find genes associated with subtypes.
- Gene Interactions.
  - Graphical models.

Breakout Discussion

- How will you use Unsupervised Learning?
- What type of big data do you work with?
- What do you hope to learn from this course?

This Course

Day 1:

1. Lecture 1: 8-8:50am - Intro
2. Lecture 2: 9-9:50am - Dimension Reduction I
3. Lecture 3: 10-10:50am - Dimension Reduction II

Break

4. Lecture 4: 11:30-12:20pm - Dimension Reduction III / Lab
5. Lecture 5: 12:30-1:20pm - Dimension Reduction Lab
6. Lecture 6: 1:30-2:20pm - Clustering I

**All times Pacific.*

This Course

Day 2:

1. Lecture 1: 8-8:50am - Clustering II
2. Lecture 2: 9-9:50am - Clustering III / Lab
3. Lecture 3: 10-10:50am - Clustering Lab

Break

4. Lecture 4: 11:30-12:20pm - Testing
5. Lecture 5: 12:30-1:20pm - Graphical Models I
6. Lecture 6: 1:30-2:20pm - Graphical Models II + Intro Final Lab

**All times Pacific.*

This Course

Day 3:

1. Lecture 1: 8-8:50am - Validation + Final Lab
2. Lecture 2: 9-9:50am - Final Lab
3. Lecture 3: 10-10:50am - Final Lab Results + Best Practices

**All times Pacific.*